Aviation Human Factors Division
Institute of Aviation

University of Illinois
at Urbana-Champaign
1 Airport Road
Savoy, Illinois 61874

A Systems Perspective on Situation Awareness II: Experimental Evaluation of a Modeling & Measurement Technique

Richard Strauss (FatWire Software)
and
Alex Kirlik (University of Illinois)

Technical Report AHFD-03-13/NTSC-03-3

May 2003

Prepared for

Naval Training Systems Center
Orlando, FL

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAY 2003		**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
A Systems Perspective on Situation Awareness II: Experimental Evaluation of a Modeling & Measurement Technique	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Training Systems Center,Orlando,FL,32813**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

We present an empirical evaluation of the utility of a systems perspective on measuring and modeling Situation Awareness (SA) in a laboratory simulation requiring submarine stealth judgments to be made in an uncertain task environment. Applying the model to a comparison of baseline versus perceptually augmented interface conditions revealed that augmentation had both positive and negative effects on SA (improving the consistency with which humans perceptually acquired information, while also increasing regression bias, suggesting that supporting reliable cue perception was accompanied by overly severe assessments made on the basis of these cues). The model was also used as the basis for a post-hoc diagnosis of the factors discriminating high and low performers. These factors were both the consistency of cue perception and the ability to consistently apply task knowledge, rather than the task knowledge per se. These findings help to verify the utility of a systems-oriented approach to measuring and modeling SA in interface-mediated, uncertain environments.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**22**	

A Systems Perspective on Situation Awareness II:

Experimental Evaluation of a Modeling & Measurement Technique

RICHARD STRAUSS

FatWire Software, New York, NY USA

ALEX KIRLIK

Institute of Aviation, Departments of Psychology, Mechanical & Industrial

Engineering, and the Beckman Institute for Advanced Science & Technology

University of Illinois at Urbana-Champaign, Savoy, IL USA

ABSTRACT

We present an empirical evaluation of the utility of a systems perspective on measuring and modeling Situation Awareness (SA) in a laboratory simulation requiring submarine stealth judgments to be made in an uncertain task environment. Applying the model to a comparison of baseline versus perceptually augmented interface conditions revealed that augmentation had both positive and negative effects on SA (improving the consistency with which humans perceptually acquired information, while also increasing regression bias, suggesting that supporting reliable cue perception was accompanied by overly severe assessments made on the basis of these cues). The model was also used as the basis for a post-hoc diagnosis of the factors discriminating high and low performers. These factors were both the consistency of cue perception and the ability to consistently apply task knowledge, rather than the task knowledge per se. These findings help to verify the utility of a systems-oriented approach to measuring and modeling SA in interface-mediated, uncertain environments.

INTRODUCTION

In a companion article (Kirlik and Strauss, 2003), we presented a systems-oriented approach to modeling both the cognitive and environmental contributors to situation awareness (SA) in interface-mediated, uncertain tasks. The technique is based on early work in the psychology judgment, and more recent efforts, particularly in the weather forecasting literature, to create increasingly diagnostic measurements of achievement in uncertain contexts. Additionally, the technique is based on even more recent developments associated with complementing cognitive modeling not only with environmental modeling, but also with formal modeling of the technological contributions to SA achievement.

As we stated in the companion article, we do not believe that this technique touches on every dimension of SA. But at the same time, it may provide an important addition to the human factors toolbox in situations where important aspects of SA involve an operator's ability to correctly infer the state or properties of a distal, uncertain environment on the basis of information available from a technological interface. To date, however, this technique has never been empirically evaluated in either psychology, forecasting, or human factors research. The goal of this article is to provide the first such evaluation of the utility of this systems, or ecological approach to SA modeling and measurement. We begin with a discussion of the laboratory task environment that served as the experimental context for this research.

THE EXPERIMENTAL TASK

The experimental simulation created for this research, called SEXTENT, was developed over a three-year period with guidance from the U.S. Navy and The Johns Hopkins Applied Physics Laboratory. In SEXTENT, the participant played the role of as a crewmember aboard a submarine, or *ownship*, and performed stealth missions (here, a mission is a single experimental trial). Stealth was defined as the ownship being undetected by any other vessels in the tactical scene. The participant's task was to quickly and accurately assess whether ownship's presence had or had not been detected, on the basis of information presented on a situation display.

Simulation Environment

Each mission was populated with objects including ownship and the three enemy vessels, or *tracks*, surrounding it. Each track was described by three characteristics: (1) Its range (or distance) from ownship; (2) Its speed; and (3) Its absolute deviation in course relative to ownship. Range and course deviation are illustrated in Figure 1. Here, a track (the circle) is shown at an indeterminate range from ownship (the square) and with three course deviations (30^o, 60^o, and 90^o). Note that. regardless of a track's bearing, (i.e., its angular position relative to ownship), its course deviation is always positive and absolute.

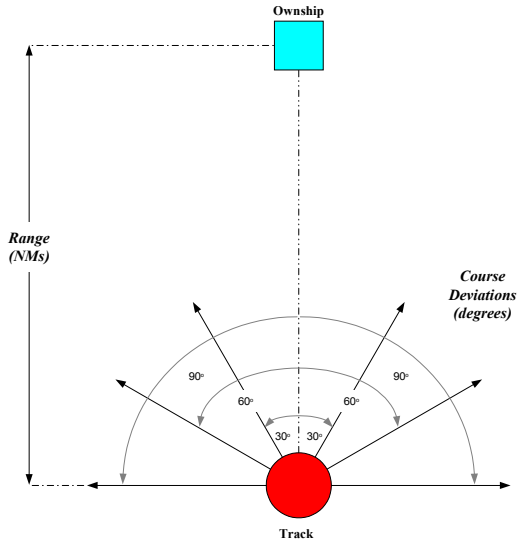

Figure 1. Geometrical Relationships Between a Track and Ownship.

Required Judgments

For each of the three tracks in a mission, a participant was requested by the "captain" of ownship to make four judgments (i.e., 12 judgments are requested per mission). The first three judgments concerned a track's characteristics. The range (R), speed (S), and course deviation (CD) were first assessed for each track. These track characteristics constitute the cues on which the participant based the fourth judgment: the probability that a track has detected ownship.

The Tactical Situation Display

The tactical situation display graphically portrayed a mission. In particular, the tactical scene graphically represented a mission's geographic boundaries, objects (i.e., tracks and ownship), and associated cues. A sample display is depicted in Figure 2. Here, ownship, shown as a square, is depicted centrally and is flanked by three tracks, shown as three circles with heading vectors.

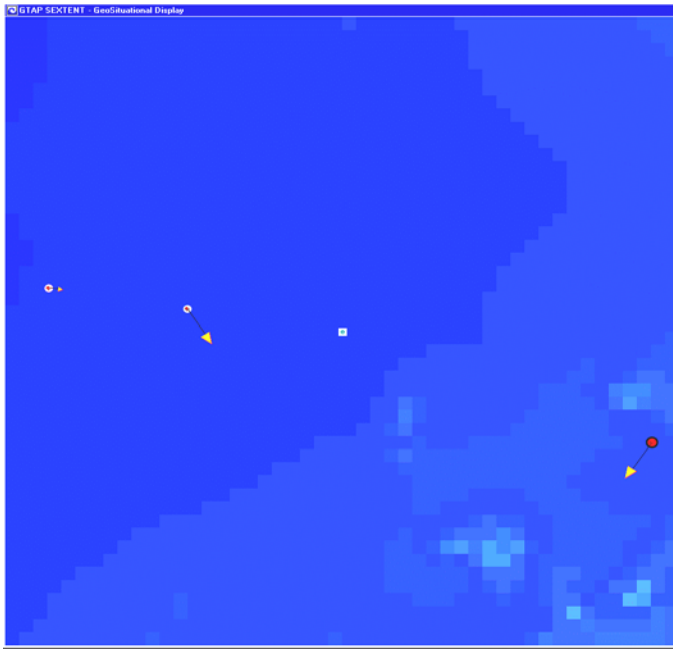

Figure 2. Baseline SEXTENT Situation Display

Track range (R) and course deviation (CD) cues are depicted on the situation display as shown in Figure 2. The speed of each track was presented as the length of the vector depicting the track's course. The longer this arrow and the wider its head, the higher the track's speed.

Situation Display Augmentation

A perceptually augmented SEXTENT display was also created. The augmentation was graphical information intended to hopefully improve the perception of a track's characteristics, thus improving SA achievement. Figure 3 shows a situation display with three forms of perceptual augmentation: (1) A range ring, (2) A speed legend, and (3) a course deviation legend. The speed and course deviation legends are shown enlarged in Figure 4. The third form of augmentation, the range ring, was intended to support judgments of range (*R*), The ring had a radius of 60 NM centered on ownship. Its right side was labeled as "60 NM."

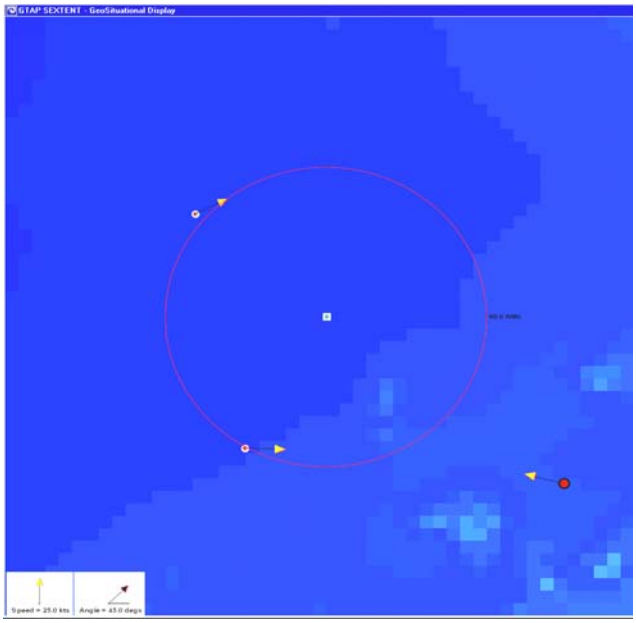

Figure 3. SEXTENT Situation Display with Perceptual Augmentation

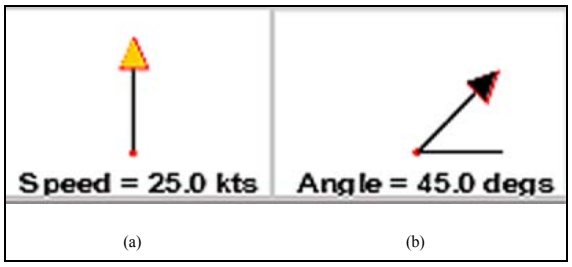

Figure 4. The Speed (a) and Course Deviation (b) Legends from the Augmented Display.

The SEXTENT Task: Formal Properties

The formal properties of the SEXTENT task environment were determined in part by sponsor input associated with representative ranges for each cue and representative directional (sign) relationships between each cue and the criterion (i.e., probability of ownship detection). This information is summarized in Table 1 and Figure 5. Tab1e 1 indicates that each of the three cues was restricted to vary over a fixed range of values. Figure 5 depicts the desired directional relationship between each cue and the criterion.

Table 1. Cue Ranges

Cue	Lower Bound	Upper Bound
R	30 NM	120 NM
S	5 kts	50 kts
CD	0 degrees	90 degrees

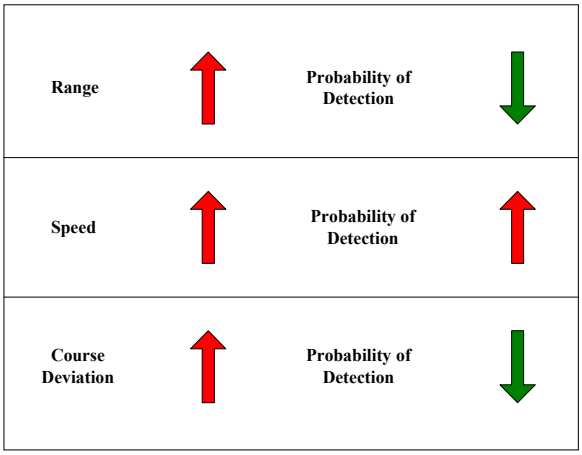

Figure 5. Directional relationships between cues and criterion.

As depicted in Figure 5, an increase in either *R* or *CD* translated into a *decrease* in the probability of detection. Conversely, an increase in *S* translated into a *increase* in the probability of detection. The design of these relationship reveals a three-way contingency—tracks with smaller ranges (i.e., closer to ownship), that maintained smaller course deviations (i.e., direct their courses toward ownship), and had higher speeds more likely had detected ownship.

Design of Cue-Criterion Relations: When designing an experimental simulation to foster the transfer of research findings to a target operational environment, one should attempt to mimic the entire set of available cues, their ranges, and their cue-criterion correlations as they exist in the target context (Brunswik, 1956; Hammond and Stewart, 2001). In our case, however, input on the nature of the task was limited to the identification of the three cues described previously, and qualitative information on their sign-relationships to the criterion. This was not a limitation in this research, since the aim was to assess the utility of the systems-oriented approach to SA modeling and measurement, rather than to make discoveries about this particular, naval setting. We did, however, have to create a concrete realization of the task, and do so in a way that was consistent with the information we were provided about cue ranges and their sign-relationships.

To create a task environment preserving the desired sign relationships and cue ranges, we created a linear model which: a) preserved each individual cue's directional relationship to the criterion; and b) over the range of cue values studied, resulted in a p(Detection) very near unity

when all three cue values maximally indicated detection, and resulted in a p(Detection) very near zero when all three cue values maximally indicated a lack of detection. The resulting function is shown in Equation 1, and has an R^2 of unity. Thus when given the three cue values, the environmental model perfectly predicts the criterion (in other words, the criterion, p(Detection), is perfectly predictable from the set of three displayed cues).

$$P(\text{Detection}) = 0.7409 + (-0.0037 * R) + (0.007 * S) + (-0.0037 * CD) \qquad \text{Equation 1}$$

Uncertainty: Given the presence of uncertainty in many if not most operational contexts, including the stealth context in which this research was performed, it was necessary to include uncertainty as a component of the SEXTENT task environment. As such, Gaussian noise with a mean of 0 and a standard deviation of .08 was added to the model shown in Equation 1. The addition of this noise reduced the R^2 of the model to .8 and thus centered each calculated probability of detection in a 99% confidence interval (CI) bound by ± 0.24 percentage points (3 standard deviations). For example, the addition of noise centered a true probability of detection equal to .7 on a 99% CI with a lower and upper bound of .46 and .94, respectively. Within each SEXTENT mission, the state of each track was independently and randomly generated by uniform sampling of the cue intervals defined in Table 1.

By precluding cue intercorrelations in this manner, we hoped that participants would be induced to attend to all three cues when making their assessments of the p(Detection). Naturally, if an operational context is characterized by redundancy (i.e., non-zero cue intercorrelations), these should be preserved in simulation-based research to foster the transfer of findings to the target context. But again, the aim of this research was evaluating the utility of an SA modeling and measurement technique, so there was no constraint to create any particular cue structures.

EXPERIMENTAL DESIGN

Our primary experimental goal was to create a set of data that could be used to assess whether the systems-oriented SA modeling and measurement technique could provide useful resources for diagnosing the factors contributing to variance in SA achievement in an interface-mediated, uncertain task. Although many possible types of variance could conceivably have been studied, in this research we focused on two sources: a) predicted variance in SA due to the use of both a baseline and a perceptually augmented situation display; and b) a post-hoc analysis of variance in SA achievement between highest and lowest scoring experimental participants.

Our logic was that by analyzing the first (display-induced) source of variance, we would be in a position to determine the degree to which the modeling and measurement technique could be used to diagnose how a design intervention may influence SA. Analogously, our purpose in analyzing the second (performer-specific) source of variance was to put us in a position to assess the degree to which the technique could be used to diagnose the sources of individual differences in SA achievement in this task. If successful, this performer-specific analysis of variance in SA could have implications for the design of training interventions targeted to the specific aspects of an individual trainee's performance that may be contributing to less than successful levels of SA.

Participants

Sixteen participants, ten men and six women, were recruited from a university student population. Each participant was randomly assigned to either the Group Baseline (GBL) display condition or to the Group Display Augmentation (GDA) display condition, resulting in a total of eight participants in each display group. For their participation, 14 of these participants were given course credit. The remaining two participants were paid an hourly rate. All participants had normal or corrected-to-normal vision. As an incentive, participants were informed that a cash prize of U.S. $50.00 would be awarded to the highest performer in each display condition.

Procedure

The first day for each participant was devoted to training. Here, each participant performed ten missions, each requiring three tracks to be judged. appropriate to his or her experimental group (i.e., either GBL or GDA). For the next eight days, participants performed 20 SEXTENT missions per day, again, each mission requiring three tracks to be judged. Over this eight-day period, this design resulted in a total 160 missions and thus 1,920 judgments per participant (160 missions X 3 tracks/mission X 4 judgments/track). For each participant, the order in which the missions were performed was randomly determined. Missions differed only in terms of the random number seed used to generate track profiles according to the environmental model.

Mission Timeline: At the start of each mission, two events occurred: (1) The tactical scene and an information panel were rendered on the main display; and (2) a *Timer* began to count the number of seconds since mission start. The information panel, located just to the right of SEXTENT's situation display, is shown in Figure 6. The panel was used to notify participants when judgments were required and to provide them with a mechanism to submit their judgments, submit these judgments, the participant pressed a green *Confirm* button.

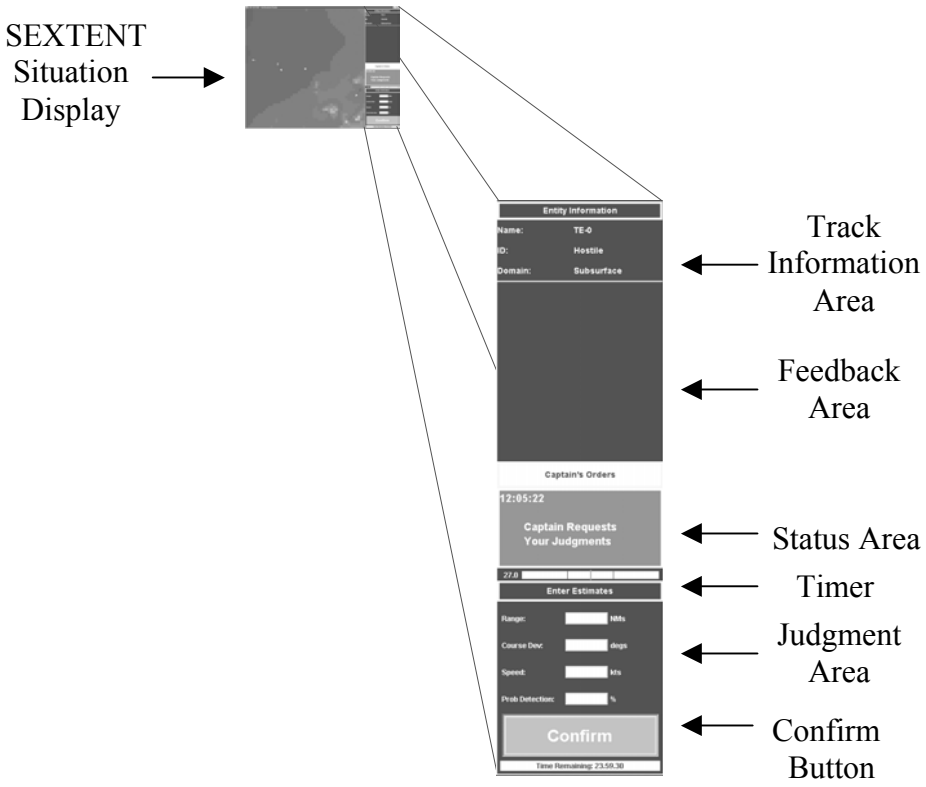

Figure 6. GT SEXTENT's Information Panel.

Three seconds after commencement, the *Status Area* in the information panel turned red and the captain's request for judgments appeared. Upon receipt of the request, a participant could begin judging any track. To initiate judgment, the participant first selected a track with a mouse. This selection had three concurrent effects. First, the selected track was highlighted in red. Second, the selected track's identifying information was posted to the *Track Information Area*. Third, the four fields in the *Judgment Area* (see Figure 6) were activated, thus turning them white and permitting the submission of judgments (prior to selection, these fields were inaccessible). After activation, the participant used the keyboard to enter the four judgments (range, speed, course deviation, and probability of detection) into the four fields. Finally, to submit these assessments, the participant pressed a green *Confirm* button.

Feedback: After judgments had been properly submitted for a track, they had been checked for illegal characters, and the "Confirm" button had been pressed, participants were given track-specific feedback. This was provided in the *Feedback* area of the information panel (see Figure 6). For *training* missions during the first day of the experiment, participants were provided with *Nine-bar Feedback*. This is shown in the *Feedback Area* of the information panel on the left side of Figure 7. Here, four pairs of bars were shown above a single bar. The top three pairs provided feedback on a participant's judgments of a track's cues (*R*, *S*, and *CD*); the fourth pair provided feedback on the judgment of a track's p(Detection), and the single bar indicated the judgment time. Each bar was suffixed with a numerical value depicting the bar's length.

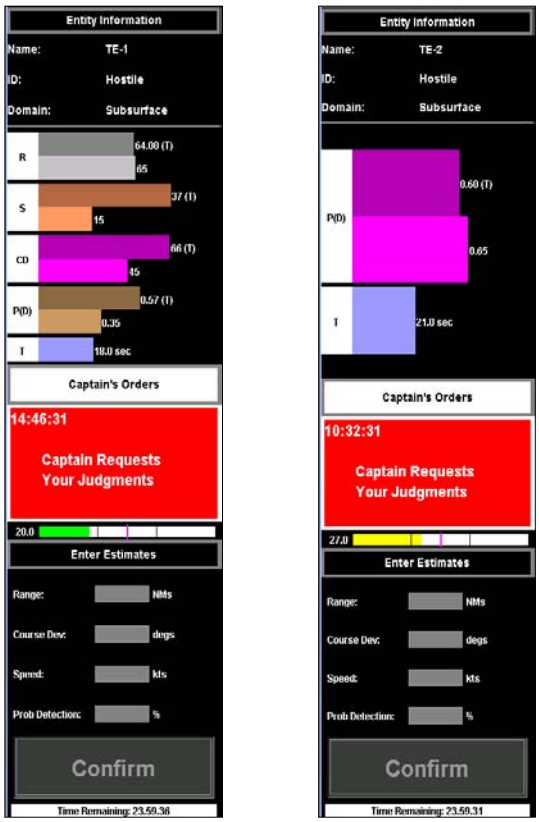

Figure 7. Nine-Bar (Day 1) and Three-Bar (Day 2-9) Feedback.

For each of the four pairs, the first bar depicted the true value of the cue or criterion being judged. The second bar depicted the participant's judgment. The "true state" bars were relatively darker in color, and their numerical values were marked with a "(T)."

For the 8 days of missions after the first training day, participants were provided with a reduced form of feedback called *Three-bar Feedback*. Three-bar feedback is depicted on the right side of Figure 7, where it can be seen to be a subset of the Nine-bar feedback used for initial training. This subset included only feedback on the submitted probability of detection (i.e., the criterion) and the time required to enter the four judgments. For more complete information on the SEXTENT task and the experimental design, see Strauss (2000).

SA MODELING AND PERFORMANCE MEASUREMENT

The systems-oriented SA measurement and modeling approach described in the companion article (Kirlik and Strauss, 2003) was used for performance modeling and measurement. In addition, knowledge of the detailed structure of the SEXTENT experimental task was used to tailor and refine the general model to provide a task-specific application of the model to the structural details particular task environment studied.

Task-Specific Application of the Modeling and Measurement Technique

Overall participant performance in SA achievement was measured using Murphy's (1988) Skill Score (SS), which, as shown in the companion article, can be decomposed as in Equation 2:

$$SS = \left[(R_{O.T})(V_{T.X})(G)(V_{U.X})(R_{Y.U}) \right]^2 - \left[r_{YO} - \left(s_Y / s_O \right) \right]^2 - \left[\left(\overline{Y} - \overline{O} \right) / s_O \right]^2 \qquad \text{Equation 2}$$

(1) (2) (3) (4) (5) (6) (7)

For this initial empirical evaluation of the SA measurement and modeling technique, we did not choose to vary the environmental factor listed as item 2 in the above equation, that is, the Fidelity of the Information System. As such, and as indicated by shading out item 2 in Table 2, we focused on only six of the seven parameters included in Equation 2 above.

Table 2. The seven components of the Expanded Lens Model

ELM Component		**Name**
	SS	Skill Score
(1)	$R_{O.T}$	Environmental Predictability
(3)	G	Knowledge
(4)	$V_{U.X}$	Consistency of Information Acquisition
(5)	$R_{Y.U}$	Consistency of Information Processing
(6)	$\left[r_{YO} - \left(s_Y / s_O \right) \right]^2$	Regression Bias
(7)	$\left[\left(\overline{Y} - \overline{O} \right) / s_O \right]^2$	Base Rate Bias

We naturally could have designed the experimental task with varying levels of technological fidelity, but this would have extended the scope of this work beyond merely the study of SA and display design, and also into the realm of human-automation interaction (HAI) (Parasuraman, Sheridan and Wickens, 2000). Hopefully, though, Equation 2 makes clear how technological fidelity, or "Stage 2 Automation," contributes to the theoretically attainable levels of SA.

Due to focusing the study in this manner, the way in which our "reduced," task-specific model of SA differs from the full model in the companion article can be readily seen in Figure 8.

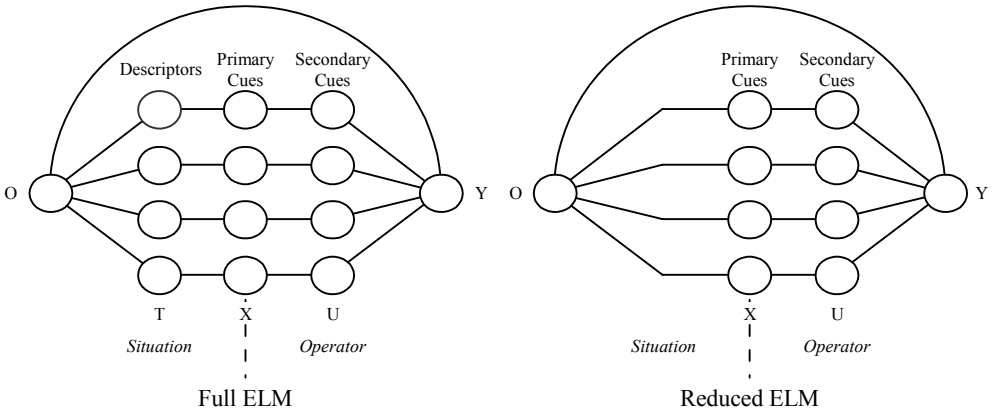

Figure 8. The original and reduced, SEXTENT-specific models of SA achievement.

In other cases, however, we were able to exploit knowledge of the experimental task to create *additional* measures representing contributions to overall SA achievement. In addition to the seven dependent variables listed in Table 2, five supplemental, model-derived, performance measures were computed, as shown in Table 3.

Table 3. Supplemental Dependent Variables

Additional Measures		**Name**
(8)	r_{YO}	Achievement
(9)	$R_{O.U}$	Secondary Environmental Predictability
(10)	V_{UX} Range	Consistency of Cue Acquisition (Range)
(11)	V_{UX} Course Deviation	Consistency of Cue Acquisition (Course Deviation)
(12)	V_{UX} Speed	Consistency of Cue Acquisition (Speed)

In Table 3, *Achievement* (8) is the traditional lens model achievement measure based on the correlation between a participant's judgments and the environmental criterion. Moreover, recall that this measure insensitive to differences in both magnitude and scale. However, when compared to the value of the more sensitive Skill Score (SS) measure, it does provide a quick way to gauge the relative contributions of both base rate bias and regression bias to overall SA achievement. *Secondary Environmental Predictability* (9) is a measure fashioned after *Environmental Predictability* (1). Yet unlike (1), which results from correlating the primary cues with the criterion, *Secondary Environmental Predictability* results from correlating the secondary cues, or participant's judgments of the cue values, with the criterion. Measure 9 thus provides a measure of the environment's predictability based on the participants *perception* of the situation, rather than based on the situational cues themselves. Its basis in the ELM is depicted in Figure 9.

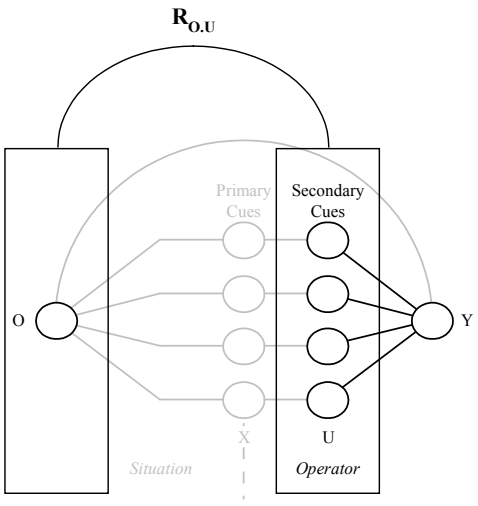

Figure 9. The grounding of *Secondary Environmental Predictability* ($R_{O.U}$).

Measures (10), (11), and (12) in Table 3 were fashioned after *Consistency of Information Acquisition*, or $V_{U.X}$ (4, from Table 2). Recall that $V_{U.X}$ measures the degree to which a correlation between a participant's judgments of the cues (i.e., the secondary cues) and criterion corresponds to a correlation between the primary cues and a participant's judgments of the criterion. The mathematical form of $V_{U.X}$ is shown in Equation 3.

$$R_{Y.X} = R_{Y.U}\left(R_{Y.X} \Big/ R_{Y.U}\right) = R_{Y.U} V_{U.X} \qquad \text{Equation 3}$$

Equation 3 depicts $V_{U.X}$ as the ratio of *Consistency* ($R_{Y.X}$) to *Consistency of Information Processing* ($R_{Y.U}$). Here, the performance of a participant whose judgments of the criterion were better predicted by the secondary than the primary cues would translate to a $R_{Y.U}$ that was greater than $R_{Y.X}$ and thus a $V_{U.X}$ that was less than one. For example, if the correlation judgments of the experimental cues (*R*, *S*, and *CD*) and judgments of the probability of detection was .9 ($R_{Y.U}$), yet the correlation between the true values of the cues and judgments of the probability of detection was .75 ($R_{Y.X}$), then $R_{Y.U}$ would be less than $R_{Y.X}$ and $V_{U.X}$ would be less than unity.

When multiple cues exist, the calculations underlying $V_{U.X}$ are based on multiple correlations, and thus provide a relatively gross perspective on information acquisition. In contrast, the detailed measures of *Consistency of Cue Acquisition* (measures 10-12 in Table 3) are based on single cues, and thus bivariate correlations. By measuring these simpler relationships, the consistency with which a *single* cue is perceived can be measured, and thus the contributions of both a primary and a secondary cue to SA achievement can be diagnosed. Note that none of the supplemental measures is articulated in the original model of SA achievement. The resources for creating these measures, however, are readily provided by the structure of that model, and thus indicate how the general framework for SA modeling and measurement can be extended and tailored to examine task-specific aspects of performance in particular environmental situations.

Statistical Analysis, Parameter Estimation, and Hypothesis Testing

Other than the Skill Score (SS) measure of overall SA achievement, every one of the measures listed in Tables 2 and 3 are correlations, and were thus adjusted for normality using Fisher's r to z_r transformation to support statistical analysis and testing (see Cooksey, 1996). All such measures were transformed back to r values for reporting in the following. MANOVA techniques reflecting the nested-factorial design (participants nested within display group) of the experiment were the primary basis for hypothesis testing, using MANOVA-integrated techniques for adjusting alpha levels as a function of the number of hypotheses tested (e.g., the multiple performance measures), as opposed to using an separate adjustment such a Bonferroni technique. However, the use of MANOVA techniques were supplemented with non-parametric testing where graphing indicated that transformed statistics clearly did not display normality.

EXPERIMENTAL RESULTS

The first day of training, which occurred on a Monday, was excluded from the analysis. The following four days of experimentation occurred during the following Tuesday through Friday, and the remaining four days of experimentation occurred the following Monday through Thursday. Since our goal was not to examine learning in the task, but instead stable performance after participants had learned the task, we present data from only the final four sessions of data collection. Readers interested in the complete set of analyses over all sessions should see Strauss (2000). In addition, we focus our presentation of results solely on the overall measure of SA achievement (SS), and those measures indicating statistically significant differences between either the two display groups, or between the two, highest and lowest scoring participants.

Results Indicating Learning, and Stability Over the Final Four Sessions: Testing indicated that over the final four days of the experiment, there was no effect of either day (block), or block X display group, on SA achievement as measured by SS. This finding supported the conclusion that participants had achieved stable levels of performance by this stage of the experiment. This result was clearly not due to the fact that participants did not learn to perform the task, as the mean SS of 0.3531 over both display groups in even the *initial* four experimental sessions was significantly greater than zero (H_0: SS 0; H_1: SS > 0; T_{127} 17.86, $p < .0001$), where zero SS indicates chance performance. By the *final* four sessions of the experiment, this mean SS across the two display groups had risen from 0.3531 to .4432, supporting our decision to focus on only the final four sessions, as participants were still learning the task up to that point. This latter finding provides additional evidence that participants indeed benefited from task experience.

Baseline Versus Augmented Displays: Somewhat surprisingly, we found no significant difference between the GBL and GDA groups in terms of overall SA achievement as measured by SS over the final four experimental sessions (Mean GBL SS = 0.4545; Mean GDA SS = .4319; $F(1,14) = 0.1945$, $p = 0.6659$). Based solely on this molar performance measure, it would be natural to conclude that perceptual augmentation had no influence on SA performance in this task. However, decomposing this molar measure of performance using the systems-oriented modeling and measurement technique resulted in a quite different set of conclusions.

A detailed analysis of display group differences indicated that the perceptually augmented display actually significantly increased some aspects of performance, while significantly

decreasing another aspect, resulting in an canceling-out effect on overall SA achievement.

Specifically, we found a significant difference between display groups on the measure of Secondary Environmental Predictability, or $R_{O.U}$ (Table 3, Measure 9). Specifically, Mean GBL $R_{O.U} = 0.856$; Mean GDA $R_{O.U} = .876$; $F(1,14) = 6.09$, $p = .027$. This difference suggests that the availability of perceptual augmentation benefited the GDA group modestly, but reliably. In particular, augmentation resulted in increased correlations between GDA participants perceptions of the environmental cues with the actual task criterion. Said another way, given the task of predicting the criterion, i.e., p[Detection), one would be significantly more accurate basing this assessment on the GDA participants' perceptions of the cue values, than on the GBL participants perceptions of the cue values. In short, the GDA participants' perceptions of the cues were more veridical and informative than GBL participants' perceptions. A submarine captain, for example, would likely do better by heeding the situational assessments (based on cue perception) of the GDA participants, rather than the situational assessments of the GBL participants, in making his or her own subsequent judgment of the probability that his or her submarine had been detected.

Why, then, did the GDA participants not outperform the GBL participants on the SS measure of SA achievement? Interestingly, the modeling technique indicated that on one measure, the GBL participants outperformed the GDA participants. Specifically, when examining the data supporting the calculation of Regression Bias (Table 2, Measure 6), we found that these data were clearly inconsistent with the normality assumption underlying parametric testing. We therefore performed a non-parametric Kruskal-Wallis test, indicating that the GBL participants had a significantly lower Regression Bias than the GDA participants (GBL Regression Bias median = .025; GDA Regression Bias median = .037; $H(1) = 5.02$, p 0.025). Recall that a regression bias manifests itself as a distribution of judgments of the task criterion over either too narrow or too broad a range (standard deviation), as compared to the standard deviation of the criterion distribution itself. These data indicate that the GDA participants displayed a regression bias [(.037 - .025)/.025] = 48% greater than did GBL participants. GDA participants rendered more severe p(Detection) estimates, in both high and low directions, than the GBL participants.

Statistically-speaking, the benefits and costs of the augmented display cancelled out, resulting in no overall difference in SA achievement. However tempting it may be, we clearly cannot conclude there is a definite causal link between the observed benefits and costs associated with display augmentation. We note, however, that other researchers have reported that display interventions using attentional cueing do improve the processing of cued information at the expense of uncued information (e.g., Yeh, Wickens and Seagull, 1999; Yeh and Wickens, 2001). In the present case, the "uncued" information was not other displayed information, but rather experientially-acquired knowledge of the distribution of actual criterion values characterizing the task ecology, information provided to participants after each trial on the feedback display.

Diagnosing the Differences Between High and Low Performers

In nearly all the statistical tests performed on both overall SA achievement (SS), and the set of model-derived performance measures, the "participant" factor was found to be significant. For example, in the four sessions on which we focused our analysis, MANOVA revealed a highly significant effect of SS across all 16 participants ($F(14,42) = 3.08$, $p = .0024$). Such findings are typical in studies of human judgment in a wide variety of uncertain task domains, prompting Brehmer and Brehmer (1988) to conclude that "All studies of policy capturing [judgment

modeling and measurement] demonstrate there are wide individual differences . . . ." (p. 103).

Given this general finding, supported by our own experimental results, we decided to see if the systems-oriented approach to SA modeling and measurement could be used, in a post-hoc manner, to shed light on the factors that may have distinguished the highest and lowest scoring performers in the SEXTENT task. We note that selecting these two performers in a post-hoc fashion is not consistent with the assumptions of standard hypothesis testing for the purpose of making *generalizable* claims about differences between *classes* of "high performers" and "low performers" in the SEXTENT task. As such, we explicitly do not intend the following results to be interpreted in this manner.

To minimize any potential for misunderstanding on this point, in the following analysis, we report statistical ratios but no explicit *p* values, which might prompt the reader to assume we intend our results to represent general conclusions about "high performers" versus "low performers" in SEXTENT. On the contrary, the following analysis had the goal of trying to understand the possible reasons why *these two* particular (high and low) performers may have differed in overall SA achievement. Although limited in this sense, we do believe that modeling and measurement tools supporting such analyses can play an important role, especially when creating performer-specific hypotheses about the particular elements of an overall task that should be targeted by individualized, performer-specific, training interventions.

Participant Differences in Overall SA Achievement: Both high and low scoring participants, who we will refer to as PH and PL respectively, came from the GBL display group. The mean SS values for PH and PL were 0.594 and 0.255, which resulted in an $F(1,18)$ ratio of 18.49. Since SS is grounded in the absolute Euclidean difference between observed data sets, we can consider SS to be on a ratio-type scale, as a zero value for SS has a non-arbitrary meaning (SS = 0 implies chance performance). As such, it is reasonable in this context to say that PH performed the task at least twice as well (0.594 versus 0.255) as PL, a striking difference in overall SA achievement.

Diagnosing the Sources of High and Low SA Achievement: One can represent the model presented in Equation 2 and the performance measures presented in Tables 2 and 3 in pictorial form, indicating the successive decomposition of the SS measure into its contributing components. Such a pictorial representation is shown in Figure 10.

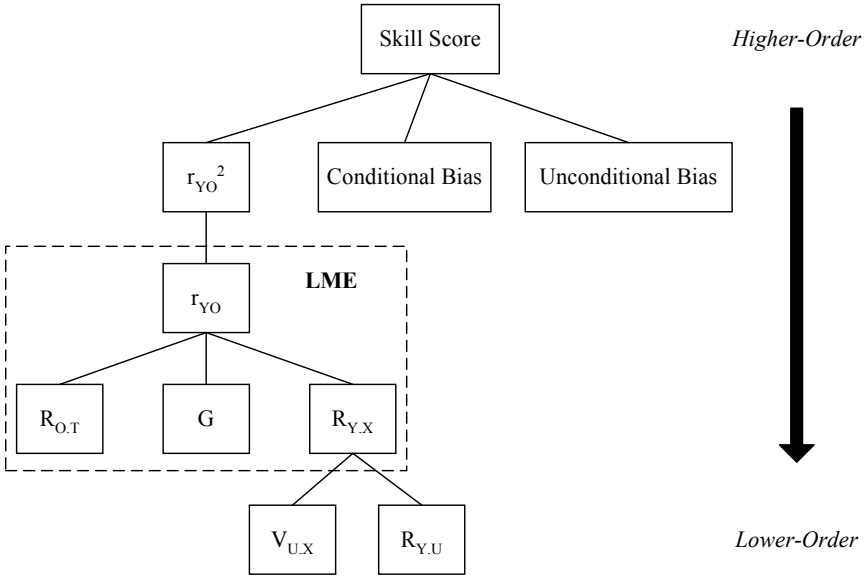

Figure 10. Pictorial Depiction of SS Decomposition

The dotted box labeled "**LME**" indicates the scope of the traditional lens model equation. To keep the complexity of the figure, and the scope of the analysis, manageable, we have omitted the manner in which the $V_{U.X}$ measure of Consistency of Cue Acquisition (see Table 2) can additionally be decomposed into the three, cue-specific acquisition measures.

Figure 10 provides a conceptual schema depicting the manner in which we proceeded to diagnose the contributors to the difference between SA achievement displayed by PH and PL. We already knew that these two performs differed strongly in terms of the top element in Figure 10, namely, in SS. Our task was then to move down the diagram shown in Figure 10 to diagnose what factors may have contributed to this molar difference.

Readers interested in the full details of the analysis should see Strauss (2000). Here, for brevity, we simply summarize the conclusions of the analysis, in pictorial form akin to that presented in Figure 10. Figure 11 shows the results of the analysis by bolding the performance measures on which statistical F ratios indicated that PH and PL differed.

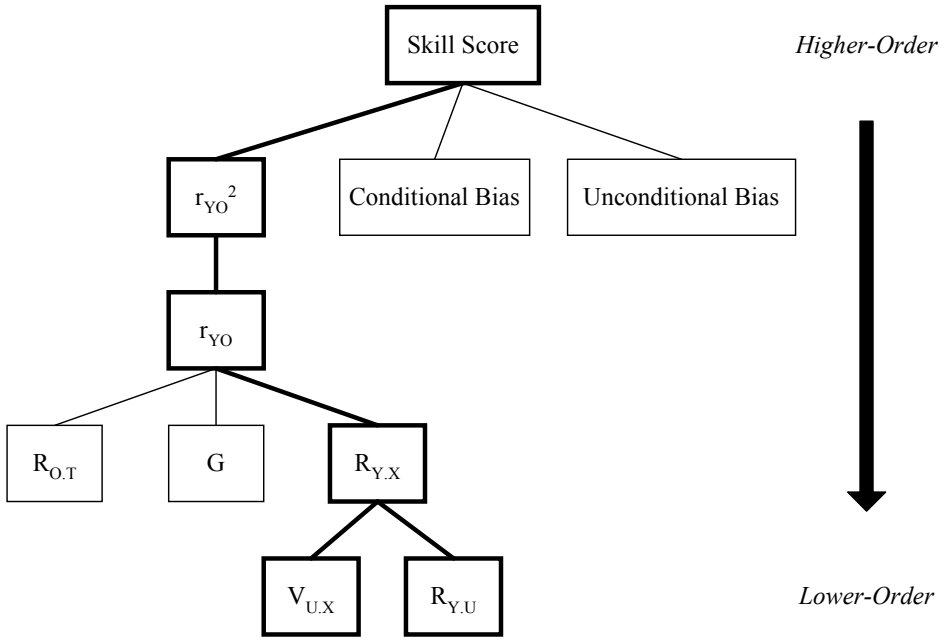

Figure 11. Results of the Analysis of Differences Between High and Low Performers

Note from Figure 11 that we were initially able to rule out both Conditional (Regression) and Unconditional (Base-Rate) biases as a potential source of SA achievement differences between PH and PL. This focused our attention on the correlation-based, lens model measure of achievement, or r_{YO}. (Mean PH r_{YO} = .816; Mean PL r_{YO} = .605; F(1,18) = 28.3). At this point, we found no differences between PH and PL on the Environmental Predictability measure ($R_{O.T}$), nor on the Task Knowledge or cue-weighting measure (G, which was very high for both PH and PL, .985 and .964 respectively). However, we did find a difference between PH and PL in terms of Consistency of Information Processing, or $R_{Y,X}$ (Mean PH $R_{Y,X}$ = .925; Mean PL $R_{Y,X}$ = .727; F(1,18) = 56.07). As shown in Figure 11, we were able to attribute this difference to both a relatively moderate effect of Consistency of Information Acquisition, or $V_{U,X}$ (Mean PH $V_{U,X}$ = 1.00; Mean PL $V_{U,X}$ .950; F(1,18) = 8.93) and a relatively larger effect of Consistency of Information Processing, or $R_{Y,U}$ (Mean PH $R_{Y,U}$ = .927; Mean PL $R_{Y,U}$ = .766; F(1,18) = 30.66).

In summary, the systems-oriented approach to SA modeling and measurement allowed us to diagnose and isolate the relatively large, two-to-one, difference in SA achievement between PH and PL to primarily the higher consistency with which PH was able to assess situations based on knowledge of the regularities of the task environment (cue weighting patterns), and to a lesser extent, the higher consistency in perceptually assessing environmental cues. Notably, this finding is consistent with that of Bisantz, Kirlik, Gay, Phipps, Walker and Fisk (2000), who similarly found that the primary difference between high and low scoring participants in their naval, combat information center, tactical judgment task was due to the ability of participants to *consistently execute* their judgment strategies in accord with task knowledge under time stress, rather than to differing levels of task knowledge itself.

CONCLUSIONS

In this article we have presented the results of modeling and measuring a phenomenon we believe to lie at the crux of situation awareness in a wide variety task situations of interest to human factors and cognitive engineering: human judgment under uncertainty in conditions where judgment is mediated by a technological interface. We have demonstrated the utility of a systems, or ecological approach to modeling SA in such contexts, by showing that the model and its associated measures provided a highly diagnostic tool for isolating the effects on SA achievement owing to both interface design interventions (a perceptually augmented situation display) and also individual differences (between high and low performers in the experimental task). In addition, we have shown that the results of these two applications of the modeling and measurement framework complement what little is already known about these phenomena.

In closing, we would like to address an issue which experience has taught us can potentially limit the acceptance, and thus the impact, of any modeling approach following in the historical tradition of Brunswik's probabilistic functionalism (Brunswik, 1956). This issue is the view that the regression-based modeling of cognition and the environment underlying the present approach is not consistent with more recent theories of perception or cognition.

On this matter, it is crucial to note that once the relevant environmental cues and criterion are identified, the assumption that a human is using a linear-additive policy for cue weighting and integration, while often appropriate (Brehmer and Brehmer, 1988; Hammond and Stewart, 2001), is hardly *required* by a systems-oriented, ecological perspective, nor does it rule out other theoretical accounts of SA. For example, to the extent that SA is relatively "direct" in the sense of Gibson (1979), and perhaps keyed tightly to a single cue or "invariant" (e.g., Smith and Hancock, 1995), this fact will fall out of regression modeling, given that the researcher has done a good job at identifying all the available sources of perceptual information (for an illustrative example, see Bisantz and Pritchett, in press).

Additionally, techniques other than regression can be used to model cue-criterion relations without sacrificing the benefits of adopting a systems-oriented perspective to SA measurement (Kirlik and Maruyama, in press). To illustrate, Campbell, Buff and Bolton (2000) described cue-criterion relations with fuzzy rules, Rothrock and Kirlik (in press) have modeled these relations with noncompensatory (if/then/and/or/not) rules, and Kirlik (1998) has modeled these relations using entropy-based, rather than regression-based techniques. All that is required to implement a systems approach to SA measurement is the availability of models for making predictions of the criterion and human assessments on the basis of cue information. The statistical calculations for estimating the many SA performance measures described in this article operate solely on cue values and model *outputs*, and are thus largely unaffected by the style of modeling used to describe the *process* by which perceptual cues are translated into levels of SA achievement.

With this said, we hope other human factors and cognitive engineering researchers continue to amend and extend the techniques presented here. Such advancements would provide even more useful methods for measuring, and thus understanding and supporting, situation awareness in human-technology interaction.

ACKNOWLEDGMENTS

We thank Johns Hopkins Applied Physics Lab and the Naval Air Warfare Center for support.

REFERENCES

Bisantz, A., Kirlik, A., Gay, P., Phipps, D., Walker, N. & Fisk, A.D. (2001). Modeling and analysis of a judgment task using a lens model approach. *IEEE* Transactions *on Systems, Man and Cybernetics – Part A: Systems and Humans, Vol. 30*, No. 6.

Brehmer, A. & Brehmer, B. (1988). What have we learned about human judgment from thirty years of policy capturing? In B. Brehmer and C.R.B. Joyce (Eds.), *Human Judgment: The SJT View* (pp. 75-114). Amsterdam: North-Holland.

Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments*. Berkeley, CA: University of California Press.

Campbell, G.E., Buff, W.L. & Bolton, A.E. (2000). The diagnostic utility of fuzzy system modeling for application in training systems. *Proceedings of the 44th Annual Meeting of the Human Factors and Ergonomics Society*. Santa Monica, CA.

Cooksey, R. W. (1996). *Judgment Analysis: Theory, Methods, and Applications*. San Diego, CA: Academic Press, Inc.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Hammond, K. R., and Stewart, T. (2001). *The Essential Brunswik: Beginnings, Explications, and Applications*. New York: Oxford University Press.

Kirlik, A. (1998). The ecological expert: acting to create information to guide action. *Fourth Symposium on Human Interaction with Complex Systems*. Dayton, OH: IEEE Computer Society Press. http://www.computer.org/proceedings /hics/8341/83410015abs.htm

Kirlik, A. & Maruyama, S. (in press). Human-technology interaction and music perception & performance: Toward the robust design of sociotechnical systems. *Proc. IEEE*.

Kirlik, A. & Strauss (2003). A systems perspective on situation awareness I: Conceptual framework, modeling, and quantitative measurement. Ms submitted for publication.

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review, 116*, 2417-2424.

Parasuraman, R., Sheridan, T.B., & Wickens, C.D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, & Cybernetics - Part A: Systems and Humans, 30*(3), 286-297.

Rothrock, L. & Kirlik, A. (in press). Inferring rule-based judgment strategies in dynamic tasks.

IEEE Systems, Man and Cybernetics – Part A: Systems and Humans.

Strauss, R. (2000). *A Methodology for Measuring the Judgmental Components of Situation Awareness.* Unpublished doctoral dissertation, School of Industrial & Systems Engineering, Georgia Institute of Technology, Atlanta, GA.

Smith, K., and Hancock, P. A. (1995). Situation awareness is adaptive, externally directed consciousness. *Human Factors, 37*(1), 137-148.

Yeh, M. & Wickens, C. D. (2001). Explicit and implicit display signaling in augmented reality: The effects of cue reliability, image realism, and interactivity on attention allocation and trust calibration. *Human Factors, 43*(3).

Yeh, M., Wickens, C. D. & Seagull, F. J. (1999). Target cueing in visual search: The effects of conformality and display location on the allocation of visual attention. *Human Factors, 41*(4), 524-542.